Asymptotic learning curve and renormalizable condition in statistical learning theory

Sumio Watanabe
P&I Lab., Tokyo Institute of Technology
4259 Nagatsuta, Midoriku, Yokohama, 226-8503 Japan
E-mail: swatanab @ pi . titech . ac . jp

March 17, 2010

Abstract

Bayes statistics and statistical physics have the common mathematical structure, where the log likelihood function corresponds to the random Hamiltonian. Recently, it was discovered that the asymptotic learning curves in Bayes estimation are subject to a universal law, even if the log likelihood function can not be approximated by any quadratic form. However, it is left unknown what mathematical property ensures such a universal law. In this paper, we define a renormalizable condition of the statistical estimation problem, and show that, under such a condition, the asymptotic learning curves are ensured to be subject to the universal law, even if the true distribution is unrealizable and singular for a statistical model. Also we study a nonrenormalizable case, in which the learning curves have the different asymptotic behaviors from the universal law.

1 Introduction

In recent studies, it was pointed out that Bayes statistics and statistical physics have the common mathematical structure, where the log likelihood function plays the same role as the random Hamiltonian, and the Bayes posterior distribution can be understood as the Boltzmann distribution. However, there are some differences between them. In statistical learning theory, the random Hamiltonian can not be necessarily approximated by any quadratic form because the Hessian matrix of the log likelihood function can be singular [1]. For example, artificial neural networks [2], normal mixtures[3], reduced rank regressions [4], Bayes networks [5], binomial mixtures, Boltzmann machines, and hidden Markov models are singular models.

The statistical properties of such models have been left unknown in statistics and information science, because it was difficult to analyze a singular likelihood function [1, 6]. Recently, new statistical learning theory has been established based on algebraic geometry, by which it was proved that the generalization and training errors are subject to a universal law, even if the statistical model does not satisfy the regularity condition [7, 8, 9, 10, 11, 12]. However, it is not yet clarified what mathematical properties ensure that such a universal law holds, therefore it is unknown the range of statistical problems which are subject to the universal law.

In this paper, we define a renormalizable condition of a statistical problem. The renormalizable condition requires that the variance function of the random Hamiltonian is bounded by the average one. We show that, if a statistical problem is renormalizable, then the algebraic geometrical method can be successfully applied, resulting that the learning curves are subject to the universal law. Also we show that, if it is not renormalizable, then the large fluctuation of the random Hamiltonian prevents the system from obeying to the universal law in general.

2 Bayes Learning Theory

Let $q(x)$ be a probability density function on N dimensional real Euclidean space $\mathbb{R}^N$. The training samples and the testing sample are respectively defined by random variables $X_1, X_2, ..., X_n$ and X, which are independently subject to the same probability distribution $q(x)dx$.

A statistical model is defined as a probability density function $p(x|w)$ of $x \in \mathbb{R}^N$ for a given parameter $w \in W \subset \mathbb{R}^d$, where W is a set of all parameters. In Bayes estimation, we prepare a probability density function $\varphi(w)$ on W. Although $\varphi(w)$ is called a prior distribution, it does not necessary represent an *a priori* knowledge of the parameter, in general.

For a given function $F(w)$ on W, its expectation value $\langle F(w) \rangle$ with respect to the posterior distribution is defined by

$$\langle F(w) \rangle = \frac{\displaystyle\int F(w) \prod_{i=1}^{n} p(X_i|w)^{\beta} \, \varphi(w) dw}{\displaystyle\int \prod_{i=1}^{n} p(X_i|w)^{\beta} \, \varphi(w) dw},$$

where $0 < \beta < \infty$ is the inverse temperature. The case $\beta = 1$ is most important because it corresponds to the strict Bayes estimation. The Bayes predictive distribution is defined by

$$p^*(x) = \langle p(x|w) \rangle.$$

In Bayes learning theory, the following random variables play an important role. The Bayes generalization loss B_g, the Bayes training loss B_t, the Gibbs generalization loss G_g, and the Gibbs training loss G_t are respectively defined by

$$\begin{aligned} B_g &= -E_X[\log \langle p(X|w) \rangle], & (1) \\ B_t &= -\frac{1}{n} \sum_{i=1}^{n} \log \langle p(X_i|w) \rangle, & (2) \\ G_g &= -\Big\langle E_X[\log p(X|w)] \Big\rangle, & (3) \\ G_t &= -\Big\langle \frac{1}{n} \sum_{i=1}^{n} \log p(X_i|w) \Big\rangle, & (4) \end{aligned}$$

where $E_X[\ ]$ shows the expectation value over X. Let us introduce two random variables by

$$\begin{aligned} Y_g &= E_X\Big[\langle (\log p(X|w))^2 \rangle - \langle \log p(X|w) \rangle^2 \Big], & (5) \\ Y_t &= \frac{1}{n} \sum_{i=1}^{n} \Big\{ \Big\langle (\log p(X_i|w))^2 \Big\rangle - \Big\langle \log p(X_i|w) \Big\rangle^2 \Big\}, & (6) \end{aligned}$$

where $V_g = nY_g$ and $V_t = nY_t$ are referred to as the *functional variances* [9, 10]. In this paper, we study the expectation values of these six random variables, which are called *Bayes observables*. The log loss function $L(w)$ and the entropy S are respectively defined by

$$\begin{aligned} L(w) &= -E_X[\log p(X|w)], \\ S &= -E_X[\log q(X)]. \end{aligned}$$

Note that $L(w) = S + D(q||p_w)$, where $D(q||p_w)$ is the relative entropy or Kullback-Leibler distance defined by

$$D(q||p_w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

Therefore, always $L(w) \geq S$. Moreover, $L(w) = S$ if and only if $p(x|w) = q(x)$. In this paper, we assume that there exists a parameter $w_0 \in W$ which minimizes $L(w)$,

$$L(w_0) = \min_{w \in W} L(w).$$

Note that such w_0 is not unique in general, because the map $w \mapsto p(x|w)$ is not one-to-one in general. We assume that, for an arbitrary w that satisfies $L(w) = L(w_0)$, $p(x|w)$ is the same probability density function. Let $p_0(x)$ be such a unique probability density function. For simplicity, we use notation $L_0 = -E_X[\log p_0(X)]$.

Definition. If $q(x) = p_0(x)$, then $q(x)$ is said to be *realizable* by $p(x|w)$, if otherwise it is said to be *unrealizable*.

Definition. If the set $W_0 = \{w \in W; p_0(x) = p(x|w)\}$ consists of a single point w_0 and if the Hessian matrix $J \equiv \nabla\nabla L(w_0)$ is strictly positive definite, then $q(x)$ is said to be *regular* for $p(x|w)$. If otherwise, then $q(x)$ is said to be *singular* for $p(x|w)$.

Bayes learning theory was studied in realizable and regular cases [13, 14, 15], realizable and singular cases [7, 9, 10], and unrealizable and regular cases [11]. In such cases, it was proved that there exists a universal relation between the generalization and training errors. In this paper, we mainly study unrealizable and singular cases.

3 Generating Function of Statistical Learning

The log density ratio function $f(x, w)$ and the log likelihood ratio function $H_n(w)$ are respectively defined by

$$\begin{aligned} f(x, w) &= \log \frac{p_0(x)}{p(x|w)}, \\ H_n(w) &= \frac{1}{n} \sum_{i=1}^{n} f(X_i, w), \end{aligned}$$

where $nH_n(w)$ is referred to as the random Hamiltonian. In this paper, we introduce the generating function of Bayes learning theory by

$$F_n(\alpha) = E\Big[-\log \int \exp(-\alpha f(X, w) - \beta n H_n(w))\varphi(w) dw\Big],$$

where $E[\ \ ]$ shows the expectation value over $X_1, X_2, .., X_n$ and X. Then, by the definitions eq.(1)-eq.(6) and by using the fact that $\log p_0(x)$ is a constant function of w, it immediately follows that

$$\begin{aligned} E[B_g] &= L_0 + F_n(1) - F_n(0), && (7) \\ E[B_t] &= L_0 + F_{n-1}(1 + \beta) - F_{n-1}(\beta), && (8) \\ E[G_g] &= L_0 + F_n'(0), && (9) \\ E[G_t] &= L_0 + F_{n-1}'(\beta), && (10) \\ E[Y_g] &= -F_n''(0), && (11) \\ E[Y_t] &= -F_{n-1}''(\beta). && (12) \end{aligned}$$

These equations show that $F_n(\alpha)$ determines the behaviors of average Bayes observables [7, 14, 15]. In order to analyze these values, we need assumptions.

Definition. If there exists a constant $\gamma > 0$ such that

$$\lim_{n\to\infty} \sup_{0\leq\alpha\leq 1+\beta} |F_n^{(3)}(\alpha)|n^\gamma = 0, \tag{13}$$

$$\lim_{n\to\infty} |F_n'(0) - F_{n-1}'(0)|n^\gamma = 0, \tag{14}$$

$$\lim_{n\to\infty} |F_n''(0) - F_{n-1}''(0)|n^\gamma = 0, \tag{15}$$

then the generating function is said to satisfy the *conditions of learnability* with index γ.

Let us assume that the conditions of learnability are satisfied. Then, by using Taylor expansions of $F_n(\alpha)$, $F_n'(\alpha)$, and $F_n''(\alpha)$, it follows that

$$E[B_g] = L_0 + F_n'(0) + \frac{1}{2}F_n''(0) + o(\frac{1}{n^\gamma}), \tag{16}$$

$$E[B_t] = L_0 + F_n'(0) + \frac{2\beta+1}{2}F_n''(0) + o(\frac{1}{n^\gamma}), \tag{17}$$

$$E[G_g] = L_0 + F_n'(0), \tag{18}$$

$$E[G_t] = L_0 + F_n'(0) + \beta F_n''(0) + o(\frac{1}{n^\gamma}), \tag{19}$$

$$E[Y_g] = -F_n''(0), \tag{20}$$

$$E[Y_t] = -F_n''(0) + o(\frac{1}{n^\gamma}). \tag{21}$$

Therefore, we obtain the *equations of states in statistical learning*,

$$E[B_g] = E[B_t] + \beta E[Y_t] + o(\frac{1}{n^\gamma}), \tag{22}$$

$$E[G_g] = E[G_t] + \beta E[Y_t] + o(\frac{1}{n^\gamma}). \tag{23}$$

That is to say, if the conditions of learnability are satisfied, then the equations of states hold. Minimization of both $E[B_g]$ and $E[G_g]$ is one of the main purposes of statistical estimation, however, they need the expectation value over the testing sample $E_X[\ ]$, hence they cannot be calculated directly from training samples. On the other hand, B_t, G_t, and Y_t can be calculated from only training samples without any direct information about $q(x)$. In other words, the equations of states show that $E[B_g]$ and $E[G_g]$ can be estimated from training samples, therefore $B_t+\beta Y_t$ and $G_t+\beta Y_t$ are information criteria which show how appropriate the set $(p(x|w), \varphi(w))$ is. In fact, they are equal to AIC [16] if $q(x)$ is realizable by and regular for $p(x|w)$. If $q(x)$ is unrealizable by or singular for $p(x|w)$, then AIC is not equal to the asymptotic generalization error, whereas $B_t + \beta Y_t$ and $G_t + \beta Y_t$ are. Hence they are called *widely applicable information criteria* (WAIC) [9, 10, 12].

4 Renormalizable Case

Let us define the renormalizability.

Definition. Let $W_\epsilon = \{w \in W; D(p_0||p_w) \leq \epsilon\}$. If there exist $A > 0$ and $\epsilon > 0$ such that

$$w \in W_\epsilon \Longrightarrow L(w) - L_0 \geq A\ D(p_0||p_w),$$

then the pair $(q(x), p(x|w))$ is said to be *renormalizable*. If otherwise, *nonrenormalizable*.

It is easy to show that, if $q(x)$ is regular for $p(x|w)$, then $(q(x), p(x|w))$ is renormalizable. In fact, $D(p_0||p_w)$ is smaller than some quadratic form of $w - w_0$ in the neighborhood of unique

w_0 and $L(w) - L_0$ has a positive definite Hessian matrix. Also, it is trivial to show that, if $q(x)$ is realizable by $p(x|w)$, then $(q(x), p(x|w))$ is renormalizable. In fact, since $q(x) = p_0(x)$, $L(w) - L_0 = D(p_0||p_w)$. However, if $q(x)$ is unrealizable by and singular for $p(x|w)$, then $(q(x), p(x|w))$ may be renormalizable or nonrenormalizable.

In this section, we study the renomalizable case, and show that the conditions of learnability hold with index $\gamma = 1$ and that the Bayes observables are subject to the universal law.

We assume that $L(w)$ is an analytic function of $w \in W$ and that $w \mapsto f(x, w)$ is a function-valued analytic function. Since $\int p_0(x)dx = \int p_w(x)dx = 1$,

$$D(p_0||p_w) = \int p_0(x)(f(x,w) + e^{-f(x,w)} - 1)dx.$$

There exists a constant $B > 0$ such that

$$\frac{t + e^{-t} - 1}{t^2} \geq B \quad (|t| < \epsilon).$$

By combining this inequality with the renormalizability, it follows that

$$L(w) - L_0 \geq AB \int p_0(x) f(x,w)^2 dx. \tag{24}$$

Since $L(w) - L_0$ is an analytic function, we can apply resolution of singularities [17, 19] to $L(w) - L_0$, and obtain the following result. There exist both a real d-dimensional analytic manifold $\mathcal{M}$ and a real analytic map $g : \mathcal{M} \to W$ such that, in each local coordinate of $\mathcal{M}$,

$$\begin{aligned} L(g(u)) - L_0 &= u^{2k} \equiv \prod_{j=1}^{d} u_j^{2k_j}, \\ |g'(u)|\varphi(g(u)) &= b(u)u^h \equiv b(u) \prod_{j=1}^{d} u_j^{h_j}, \end{aligned}$$

where $k = (k_1, k_2, ..., k_d)$ and $h = (h_1, h_2, ..., h_d)$ are multiple indeces made of nonnegative integers, $|g'(u)|$ is the Jacobian determinant of the map $w = g(u)$, and $b(u) > 0$. Then, by using eq.(24), $f(x, g(u))^2$ can be divided by u^{2k}, in other words, $f(x, g(u))/u^k$ is a well-defined analytic function. In fact, if $f(x, g(u))$ can not be divided by u^{2k}, then eq.(24) does not hold. Hence, there exists a function-valued analytic function $a(x, u)$ such that

$$f(x, g(u)) = a(x, u)u^k.$$

Moreover, from $L(w) - L_0 = E_X[f(X, w)]$, we have $E_X[a(X, u)] = u^k$. Remark that both renormalizability and resolution theorem are necessary to prove the existence of $a(x, u)$. Let us define an empirical process on $\mathcal{M}$,

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{a(X_i, u) - u^k\}.$$

Then the probability distribution of $\xi_n(u)$ converges to that of the gaussian process $\xi(u)$, which is uniquely determined by its average and covariance [9, 18],

$$\begin{aligned} E_\xi[\xi(u)] &= 0, \\ E_\xi[\xi(u)\xi(u')] &= E_X[a(X,u)a(X,u')] - E_X[a(X,u)]E_X[a(X,u')], \end{aligned}$$

where $E_\xi[\ ]$ shows the expectation value over the gaussian process $\xi(u)$. Moreover, the gaussian process $\xi(u)$ can be represented by

$$\xi(u) = \sum_{j=1}^{\infty} c_j(u) g_j$$

where $\{g_j\}$ are independent random variables and each g_j is subject to the standard normal distribution. Then

$$E_\xi[\xi(u)\xi(u')] = \sum_{j=1}^{\infty} c_j(u)c_j(u').$$

The random Hamiltonian is rewritten as

$$nH_n(g(u)) = nu^{2k} - \sqrt{n}u^k\xi_n(u).$$

To study the generatining function $F_n(\alpha)$, we need the asymptotic behavior of

$$Z_n(s) = \int f(x,w)^s \exp(-\beta n H_n(w))\varphi(w)dw,$$

where $s \geq 0$ is a real value. For example,

$$F_n'(0) = E\Big[\frac{Z_n(1)}{Z_n(0)}\Big], \tag{25}$$

$$F_n''(0) = -E\Big[\frac{Z_n(2)}{Z_n(0)}\Big] + E\Big[\frac{Z_n(1)}{Z_n(0)}\Big]^2. \tag{26}$$

Then by using the function $w = g(u)$,

$$\begin{aligned} Z_n(s) &= \sum_\alpha \int du\ a(x,u)^s u^{sk+h} \exp(-\beta n u^{2k} + \beta\sqrt{n}u^k\xi_n(u))b_\alpha(u) \\ &= \sum_\alpha \int_0^\infty dt \int du\ \frac{1}{n}\ \delta\Big(\frac{t}{n} - u^{2k}\Big)\ a(x,u)^s u^{sk+h} \exp(-\beta t + \beta\sqrt{t}\xi_n(u))b_\alpha(u), \end{aligned}$$

where $\sum_\alpha$ shows the sum over all local coordinates and $b_\alpha(u) \geq 0$ satisfies $\sum_\alpha b_\alpha(u) = b(u)$. By using the asymptotic expansion of the Shcwartz distribution $\delta(t/n - u^{2k})$ for $n \to \infty$ [7, 9, 10, 20, 21, 22, 25, 26], there exists a Schwartz distribution $D_\alpha(u)$ such that

$$\sum_\alpha \frac{1}{n}\ \delta\Big(\frac{t}{n} - u^{2k}\Big)\ u^{sk+h}\ b_\alpha(u) \cong \frac{(\log n)^{m-1}}{n^{\lambda+s/2}}\ t^{\lambda-1+s/2}\Big(\sum_{\alpha^*} D_{\alpha^*}(u)\Big),$$

where $\lambda > 0$ is the *log canonical threshold* defined by

$$\lambda = \min_\alpha \min_{j=1}^{d}\Big(\frac{h_j+1}{2k_j}\Big),$$

and m is the maximum number of j which attains the above minimum. Also $\sum_{\alpha^*}$ shows the sum over all local coordinates that attain the above minimum and the support of $D_{\alpha^*}(u)$ is contained in the set $\{u \in \mathcal{M}; L(g(u)) - L_0 = 0\}$. Hence

$$Z_n(s) \cong \frac{(\log n)^{m-1}}{n^{\lambda+s/2}}\Big(\int \mathcal{D}(u,t)t^{s/2}\exp(\beta\sqrt{t}\xi(u))\Big).$$

where $\int \mathcal{D}(u,t)$ is defined by the integration over the manifold,

$$\int \mathcal{D}(u,t) = \sum_{\alpha^*}\int_0^\infty dt \int du D_{\alpha^*}(u)\ t^{\lambda-1}\ \exp(-\beta t).$$

Let us define

$$\hat{Z}(q,r,s) = \int \mathcal{D}(u,t)\ \xi(u)^q\ t^{r/2}\ a(x,u)^s\ \exp(\beta\sqrt{t}\xi(u)).$$

Then

$$Z_n(s) \cong \frac{(\log n)^{m-1}}{n^{\lambda+s/2}} \hat{Z}(0,s,s). \tag{27}$$

Firstly, since $E_X[a(X,u)] = u^k$,

$$E_X[\hat{Z}(0,1,1)] = \hat{Z}(0,2,0).$$

Secondly, by using the partial integration of t

$$\int_0^\infty dt\, t^\lambda e^{-\beta t+\beta\sqrt{t}\xi(u)} = \frac{\lambda}{\beta}\int_0^\infty dt\, t^{\lambda-1}e^{-\beta t+\beta\sqrt{t}\xi(u)} + \frac{1}{2}\int_0^\infty dt\, t^{\lambda-1/2}\xi(u)e^{-\beta t+\beta\sqrt{t}\xi(u)},$$

it follows that

$$\hat{Z}(0,2,0) = \frac{\lambda}{\beta}\hat{Z}(0,0,0) + \frac{1}{2}\hat{Z}(1,1,0).$$

And lastly, by using the partial integration over the gaussian process $\xi(u)$,

$$\begin{aligned}
E_\xi\Big[\frac{\hat{Z}(1,1,0)}{\hat{Z}(0,0,0)}\Big] &= E_\xi\Big[\int \mathcal{D}(u,t)\Big(\sum_{j=1}^\infty c_j(u)g_j\Big)\frac{t^{1/2}\exp(\beta\sqrt{t}\xi(u))}{\int \mathcal{D}(u',t')\exp(\beta\sqrt{t'}\xi(u'))}\Big] \\
&= E_\xi\Big[\int \mathcal{D}(u,t)\Big(\sum_{j=1}^\infty c_j(u)\frac{\partial}{\partial g_j}\Big)\frac{t^{1/2}\exp(\beta\sqrt{t}\xi(u))}{\int \mathcal{D}(u',t')\exp(\beta\sqrt{t'}\xi(u'))}\Big] \\
&= \beta E_X E_\xi\Big[\frac{\hat{Z}(0,2,2)}{\hat{Z}(0,0,0)}\Big] - \beta E_X E_\xi\Big[\frac{\hat{Z}(0,1,1)}{\hat{Z}(0,0,0)}\Big]^2,
\end{aligned} \tag{28}$$

where we used $E_\xi[\xi(u)\xi(u')] = E_X[a(X,u)a(X,u')]$ on the set $\{u; L(g(u)) - L_0 = 0\}$. Let us define the constant 2ν by the right hand side of eq.(28), where ν is referred to as the *singular fluctuation*. Then by using eqs.(25),(26),(27),

$$\begin{aligned}
F_n'(0) &\cong (\frac{\lambda}{\beta}+\nu)\cdot\frac{1}{n}, \\
F_n''(0) &\cong -\frac{2\nu}{\beta}\cdot\frac{1}{n}.
\end{aligned}$$

Therefore, we obtained the *universal law* of Bayes observables,

$$E[B_g] \cong L_0 + (\frac{\lambda-\nu}{\beta}+\nu)\frac{1}{n}, \tag{29}$$

$$E[B_t] \cong L_0 + (\frac{\lambda-\nu}{\beta}-\nu)\frac{1}{n}, \tag{30}$$

$$E[G_g] \cong L_0 + (\frac{\lambda}{\beta}+\nu)\frac{1}{n}, \tag{31}$$

$$E[G_t] \cong L_0 + (\frac{\lambda}{\beta}-\nu)\frac{1}{n}, \tag{32}$$

$$E[Y_g] \cong E[Y_t] \cong \frac{2\nu}{\beta}\cdot\frac{1}{n}. \tag{33}$$

In this case, we can prove that the conditions of learnability with index $\gamma = 1$ are satisfied by the same way as [9, 10]. Hence, equations of states hold with $\gamma = 1$.

5 Nonrenormalizable Case

In this section, we study a nonrenormalizable case. It is still difficult to clarify the general nonrenormalizable case. Hence, in this section, we show that there exists a simple example in which the Bayes observables do not satisfy the universal law.

$$q(x,y) = \frac{1}{2\pi}\exp(-\frac{1}{2}(x^2+y^2)), \tag{34}$$

$$p(x,y|a) = \frac{1}{2\pi}\exp(-\frac{1}{2}\{(x-a)^2+(y-\sqrt{a^4-a^2+1})^2\}), \tag{35}$$

where $a \in \mathbb{R}^1$ is a parameter. Then the relative entropy is

$$D(q||p_a) = \int q(x,y)\log\frac{q(x,y)}{p(x,y|a)}dxdy = \frac{1}{2}(a^4+1).$$

Hence $D(q||p_a)$ is minimized at $a=0$, and $L_0 = \log(2\pi)+3/2$. The Hessian is given by $\partial_a^2 D(q||p_a)|_{a=0} = 0$. Therefore $q(x)$ is unrealizable by and singular for $p(x|a)$. The log density ratio function is

$$f(x,a) = -ax - h(a)y + \frac{a^4}{2},$$

where $h(a) = \sqrt{a^4-a^2+1}-1$ is a real analytic function, and

$$D(p_0||p_a) = \frac{a^4}{2} - h(a).$$

Note that $D(p_0||p_a) \cong a^2/2$ in the neighborhood of $a=0$. On the other hand, $L(a)-L_0 = a^4/2$, resulting that $(q(x), p(x|a))$ is not renormalizable. The random Hamiltonian is

$$nH_n(a) = \frac{n\,a^4}{2} - \sqrt{n}\,a\,\xi_1 - \sqrt{n}\,h(a)\,\xi_2,$$

where

$$\xi_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i, \quad \xi_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$$

are independently subject to the standard normal distribution.

$$\begin{aligned} nH_n(a)' &= 2\,n\,a^3 - \sqrt{n}\,\xi_1 - \sqrt{n}\,h(a)'\,\xi_2, \\ nH_n(a)'' &= 6\,n\,a^2 - \sqrt{n}\,h(a)''\,\xi_2. \end{aligned}$$

The parameter a that minimizes $nH_n(a)'$ is denoted by a^*. Since

$$a^* = \Big(\frac{\xi_1}{kn}\Big)^{1/3} + o_p(\frac{1}{n^{1/3}}),$$

the main order term of $nH_n(a)$ is given by

$$\begin{aligned} nH_n(a) &= \frac{1}{2}nH_n(a^*)(a-a^*)^2 + nH_n(a^*) \\ &= \frac{1}{2}C_n(a-D_n)^2 - \frac{1}{2}C_nD_n^2, \end{aligned}$$

where

$$\begin{aligned} C_n &= 6n(\xi_1/2\sqrt{n})^{2/3}, \\ D_n &= (\xi_1/2\sqrt{n})^{1/3}. \end{aligned}$$

Therefore, by using $E[(\xi_1)^{\mu-1}] = 2^{\mu/2}\Gamma(\mu/2)/\sqrt{2\pi}$,

$$\begin{aligned} F_n'(0) &= \frac{Q}{2} \cdot \frac{1}{n^{2/3}}, \\ F_n''(0) &= -\frac{2Q}{\beta} \cdot \frac{1}{n^{2/3}}, \end{aligned}$$

where

$$Q = \frac{2^{7/6}}{\sqrt{2\pi}}\Gamma(\frac{7}{6}).$$

The asymptotic behaviors of Bayes observables are different from the universal law,

$$E[B_g] \cong L_0 + (\frac{1}{2} - \frac{1}{\beta}) \cdot \frac{Q}{n^{2/3}}, \tag{36}$$

$$E[B_t] \cong L_0 - (\frac{3}{2} + \frac{1}{\beta}) \cdot \frac{Q}{n^{2/3}}, \tag{37}$$

$$E[G_g] \cong L_0 + \frac{Q}{2} \cdot \frac{1}{n^{2/3}}, \tag{38}$$

$$E[G_t] \cong L_0 - \frac{3Q}{2} \cdot \frac{1}{n^{2/3}}, \tag{39}$$

$$E[Y_g] \cong E[Y_t] \cong \frac{2Q}{\beta} \cdot \frac{1}{n^{2/3}}. \tag{40}$$

Also in this case, the conditions of learnability are satisfied with index $2/3$, hence the equations of states hold with $\gamma = 2/3$, however,

$$E[V_t] = nE[W_t] \cong n^{1/3}$$

does not converge to the constant. It seems that both renormalizable and nonrenormalizable statistical problems satisfy the more general universal law.

6 Discussion

In this section, let us discuss three points, birational invariants, renormalizability, and Bayes observables as random variables.

6.1 Birational Invariants

In section 4, we proved that, in the renormalizable case, the asymptotic learning curves are determined by λ and ν, which are defined by using resolution of singularities. Let us study the mathematical properties of them. For a given analytic function, $L(w) - L_0$, there exist infinitely many desingularization pairs $(\mathcal{M}, g)$. If a value defined by using $(\mathcal{M}, g)$ does not depend on the choice of $(\mathcal{M}, g)$, then it is called a *birational invariant*.

Firstly, as is shown in [7, 9], the value $(-\lambda)$ is equal to the largest pole of the *zeta function* on $\mathbb{C}$ obtained by the analytic continuation of

$$\zeta(z) = \int (L(w) - L_0)^z \varphi(w) dw \quad (\mathrm{Re}(z) > 0).$$

Therefore, λ is a birational invariant. This value is well known in algebraic geometry and algebraic analysis, which shows the relative relation of the pair of two algebraic varieties (W, W_0) [17, 20, 23, 24, 25, 26].

Secondly, the value ν is characterized by

$$\nu = \lim_{n\to\infty} \frac{\beta}{2} E\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{\Big\langle (\log p(X_i|w))^2 \Big\rangle - \Big\langle \log p(X_i|w) \Big\rangle^2\Big\}\Big].$$

Hence ν is also a birational invariant.

It was clarified by [11] that, if a true distribution is unrealizableby and regular for a parametric model, then

$$\begin{aligned} \lambda &= d/2, \\ \nu &= \mathrm{tr}(IJ^{-1})/2, \end{aligned}$$

where I and J are respectively $d \times d$ matrices defined by

$$\begin{aligned} I &= E_X[\nabla \log p(x|w_0) \nabla \log p(x|w_0)], \\ J &= \nabla^2 L(w_0). \end{aligned}$$

For singular and realizable cases, λ was calculated in [3, 4, 5, 10], whereas ν is unknown.

6.2 Renormalizability

Let us discuss the renormalizable condition.

Firstly, we study the renormalizable condition from the physical point of view. In physics, a set of functions $\{f_n(x); n = 1, 2, ...,\}$ is sometimes called renormalizable if there exists some rescaling transform by which a universal law is discovered. For example, if there exist both a set (a, b) and a function $f^*(x)$ such that

$$\lim_{n\to\infty} n^a f_n(n^b x) \to f^*(x),$$

then such a system is called renormalizable. In this paper, we have shown that, if $(q(x), p(x|w))$ is renormalizable, then the Boltzmann distribution satisfies the convergence in law,

$$\frac{n^\lambda}{(\log n)^{m-1}} \exp(-n\beta H_n(g(u))) \to \int_0^\infty t^{\lambda-1} \exp(-n\beta t + \beta\sqrt{t}\,\xi(u))dt,$$

when n tends to infinity, where $\xi(u)$ is a gaussian process defined by the central limit theorem of the functional space. If $(q(x), p(x|w))$ does not satisfy the renormalizable condition, then such a rescaling transform does not exist in general. The expectation and the variance of

$$nH_n(w) = \sum_{i=1}^{n} f(X_i, w)$$

are respectively given by

$$E[nH_n(w)] = nE_X[f(X, w)], \quad V[nH_n(w)] = nV_X[f(X, w)].$$

Because $E_X[f(X, w)] = L(w) - L_0 \geq 0$ and $V_X[f(X, w)] \cong (1/2)D(p_0||p_w)$ in the neighborhood $L(w) - L_0 = 0$, the renormalizable condition ensures that the fluctuation of the random Hamiltonian is bounded by the average one. This is the intuitive reason why the universal law holds.

Secondly, we study scale invariantness of renormalizablity. Let $f_1(x, w)$ and $f_2(x, w)$ be log likelihood ratio functions of two different statistical problems. If they are renormalizable and satisfy the relations

$$\begin{aligned} E_X[f_1(X, w)] &= E_X[f_2(X, w)], \\ E_X[f_1(X, w)f_1(X, w')] &= E_X[f_2(X, w)f_2(X, w')], \end{aligned}$$

then they have the same birational invariants (λ, ν). In other words, the learning curves are determined only by the average and covariance of the log density ratio function. It might seem that $E_X[f(X,w)]^2 \propto E_X[f(X,w)^2]$, but such a relation does not hold even in a trivial case. In a realizable and regular case, $a \in \mathbb{R}^1$,

$$p(x|a) = \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}(x-a)^2)$$

and $q(x) = p(x|0)$, then $f(x,a) = a^2/2 - ax$, resulting that $E_X[f(X,a)] = a^2/2$ and $E_X[f(X,a)^2] \cong a^2 + a^4/4$. Therefore, in the neighborhood of $a = 0$, both $E_X[f(X,a)]$ and $E_X[f(X,a)^2]$ are in proportion to a^2. The renormalizable condition in this case is invariant under a scaling transform $f(X,w) \to sf(X,w)$ for an arbitrary constant $s > 0$. The renormalizable condition of this paper is a generalized concept of such invariantness.

6.3 Bayes Observables as Random Variables

In statistical learning theory, Bayes observables are random variables. In this paper, we mainly studied the expectation values of them. Note that the generating function $F_n(\alpha)$ does not have sufficient information about randomness of Bayes observables. If a true distribution is regular or realizable, then stochastic properties of Bayes observables were clarified [10, 11]. It is a future study to clarify the stochastic behavior of Bayes observables as random variables.

7 Conclusion

In this paper, we defined the renormalizable condition of a learning system, and proved that, in the renormalizable case, the universal law holds. Also we showed that, in nonrenormalizable case, the universal law does not hold in general. It is the future study to clarify the more general universal learning theory, which contains both renormalizable and nonrenormalizable statistical problems.

Acknowledgment. This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

References

[1] Watanabe S 1995 *Proc. of Int. Symp. on NOLTA* 207-210

[2] Watanabe S 2001 *Neur. Netw.* **14**(8) 1049-60

[3] Yamazaki K and Watanabe S 2003 *Neur. Netw.* **16**(7) 1029-38

[4] Aoyagi M and S.Watanabe S 2005 *Neur. Netw.* **18**(7) 924-33

[5] Rusakov D and Geiger D 2005 *J. of Mach. Lear. Res.* **6** 1-35

[6] Hartigan J A 1985 Proc. *Barkeley Conference in Honor of J. Neyman and J. Kiefer* **2** 807-10

[7] Watanabe S 2001 *Neur. Comput.* **13**(4) 899-933

[8] Drton M, Sturmfes B and Sullivant S 2009 *Lectures on Algebraic Statistics* (Berlin: Birkhäuser Verlag)

[9] Watanabe S 2009 *Algebraic geometry and statistical learning theory* (Cambridge: Cambridge University Press)

[10] Watanabe S 2010 *Neur. Netw.* **23**(1) 20-34

[11] Watanabe S 2010 *IEICE Trans* **E93-A**(3) 617-26

[12] Watanabe S 2010 *Adv. Stud. Pure Math.* **57** 473-92

[13] Schwarz G 1978 *Ann. Stat.* **6**(2) 461-4

[14] Levin E, Tishby N and Solla S A 1990 *Proc. of IEEE,* **78**(10) 1568-74

[15] Amari S 1993 *Neur. Netw.* **6**(2) 161-6

[16] Akaike H 1974 *IEEE Trans. on Aut. Cont.* **10** 716-23

[17] Hironaka H 1964 *Ann. of Math.* **79** 109-326

[18] van der Vaart A W and Wellner J A 1996 *Weak Convergence and Empirical Processes* (New York: Springer)

[19] Atiyah M F 1970 *Comm. Pure and Appl. Math.* **13** 145-50

[20] Bernstein I N 1972 *Func. Anal. Appl.* **6** 26-40

[21] Gelfand I M and Shilov G E 1964 *Generalized Functions* (San Diego: Academic Press)

[22] Kashiwara M 1976 *Invent. Math.* **38** 33-53

[23] Kollór J, Mori S, Clemens C H and Corti A 1998 *Birational geometry of algebraic varieties* (Cambridge: Cambridge University Press)

[24] Mustata M 2002 *J. Amer. Math. Soc.* **15** 599-615

[25] Oaku T 1997 *J. Pure Appl. Alg.* **117** 495-518

[26] Saito M 2007 *arXiv:0707.2308v1*